(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（A）　(11)特許出願公開番号

特開平10－95796

(43)公開日　平成10年(1998)４月14日

(51)Int.Cl.[6]	識別記号	FI	
C07K 14/00		C07K 14/00	
G01N 33/68		G01N 33/68	
// G06F 17/00		G06F 15/20	D

審査請求 有　請求項の数2　OL　（全 8 頁）

(21)出願番号　特願平9－283233

(62)分割の表示　特願平5－246805の分割

(22)出願日　平成５年(1993)10月１日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　馬見塚　拓
東京都港区芝五丁目７番１号　日本電気株式会社内

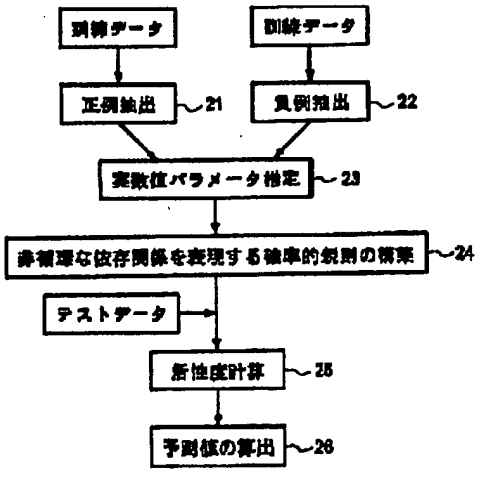
(74)代理人　弁理士　京本　直樹　（外２名）

(54)【発明の名称】　タンパク質立体構造予測方法

(57)【要約】

【課題】　構造未知のタンパク質のアミノ酸配列情報から、その立体構造の各局所構造を高精度で予測する。

【解決手段】　ステップ２１で構造既知及び未知のタンパク質アミノ酸配列を入力とし、それらのアライメント（整合）から局所構造領域の正例を出力し、ステップ２２で立体構造既知のタンパク質アミノ酸配列を入力とし、それらのアライメントから局所構造領域の負例を出力し、ステップ２３で前記正例と負例を入力とし、これら訓練データのアミノ酸種類あるいは実数値属性から確率的規則の実数値パラメータの推定値を計算し、ステップ２４では情報量規準に基づき、各残基位置に相当する変数間の依存関係を表現する確率的規則の構造を決定し、ステップ２５でテストデータ配列を入力とし、前記確率的規則を使用し、テストデータ配列の各領域が局所構造であるかの確からしさを示す活性度を出力し、ステップ２６で前記活性度を入力とし、その中から最適値を出力する。



1

【特許請求の範囲】

【請求項1】タンパク質構造の規則を学習するステップを有するタンパク質立体構造予測方法において、前記ステップが、正例と負例とからなる訓練データのアミノ酸配列を入力とし、各依存関係を条件付き確率とし、あらかじめ定められた情報量基準を用いて、該アミノ酸配列の各々の残基位置が依存している他の残基位置を、依存関係が循環しないという制限の下で決定することにより、該依存関係を表現する確率的規則を構築して出力することを特徴とするタンパク質立体構造予測方法。

【請求項2】タンパク質構造の規則を学習するステップを有するタンパク質立体構造予測方法において、前記ステップが、正例と負例とからなる訓練データのアミノ酸配列の実数値属性を入力とし、各依存関係を条件付き確率とし、あらかじめ定められた情報量基準を用いて、該アミノ酸配列の各々の残基位置が依存している他の残基位置を、依存関係が循環しないという制限の下で決定することにより、該依存関係を表現する確率的規則を構築して出力することを特徴とするタンパク質立体構造予測方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、立体構造未知のタンパク質アミノ酸配列から、タンパク質の立体構造を予測する方法に関する。

【0002】

【従来の技術】タンパク質の局所的な立体構造として、αヘリックスやβシートに代表される二次構造やジンクフィンガーやロイシンジッパーに代表されるモチーフなどがある。立体構造未知のタンパク質アミノ酸配列に対して、これらタンパク質の局所構造の予測が可能になれば、タンパク質全体の立体構造予測が可能になると一般に考えられている。

【0003】例えば、タンパク質二次構造予測問題は、20年以上も前から解決が図られてきた問題であり、従来、タンパク質の一次構造の各残基（以下、予測対象となる残基を中心残基と呼ぶ）が、3（あるいは4）種類の二次構造のいずれに対応するかを予測する問題として扱われてきた。従来技術によるタンパク質の二次構造を予測する方法として、例えば、1974年発行の米国の雑誌「バイオケミストリー」（Biochemistry）の第23巻222－245頁掲載のチョウ（Chou）とファスマン（Fasman）による論文「プレディクション オブ プロテイン コンホメーション」（Prediction of protein conformation）（以下、CF法と略す）、1978年発行の米国の雑誌「ジャーナル オブ モレキュラ バイオロジー」（Journal of Molecular Biology）の第120巻97－120頁掲載のガルニエ（Garnier）らによる論文

2

「アナリシス オブ ザ アキュレシー アンド インプリケーションズ オブ シンプル メソード フォー プレディクティング ザ セコンダリー ストラクチャー オブ グロブラープロテインズ」（Analysis of the accuracy and implications of simple method forpredicting the secondary structure of grobular proteins）（以下、GOR法と略す）、1987年発行の米国の雑誌「ジャーナル オブ モレキュラ バイオロジー」（Journal of Molecular Biology）の第198巻425－443頁掲載のギブラト（Gibrat）らによる論文「ファザー デベロプメンツ オブ プロテイン セコンダリー ストラクチャ プレディクションユージング インホメーション セオリー：ニュー パラメータズ アンド コンシダレーション オブ レジデュー ペアズ」（Further developments of protein secondary structure prediciton using information theory:New parameters and considerationof residue pairs）（以下、GGR法と略す）、1988年発行の米国の雑誌「ジャーナル オブ モレキュラ バイオロジー」（Journal of Molecular Biology）の第202巻865－884頁掲載のキャン（Qian）らによる論文「プレディクティング ザ セコンダリー ストラクチャー オブ グロブラー プロテインズ ユージング ニューラル ネットワーク モデルズ」（Predicting the secondary structure of globular protein susing neural network models）（以下QSと略す）、及び1993年の米国の学会「ハワイ インターナショナル コンファレンス オブ システム サイエンシイズ」（Hawaii International Conference of System Sciences）の予稿集第1巻659－668頁記載の馬見塚らによる論文「プロテイン αヘリックス リージョン プレディクション ベースド オン ストキャスティックルール ラーニング」（Protein α－Helix Region Prediction Based－on Stochastic Rule Learning）（以下、MY法と略す）などがある。

【0004】CF法は、タンパク質構造のデータベースから各二次構造におけるアミノ酸の統計的な出現頻度を求め、この頻度表を使用し、経験的な規則に基づく予測を行っている。また、GOR法は、中心残基の二次構造に対して、その残基から数残基離れた残基により独立に

もたらされる情報量の和を計算し、その相対値から予測を行い、ＧＧＲ法は、中心残基の二次構造に対して、その残基及びその残基から数残基離れた残基によりもたらされる情報量の和から予測を行っている。さらに、ＱＳ法は、３層のフィードフォワード型のネットワークを使用し、中心残基の前後８残基を含む配列を入力とし、二次構造に対する中心残基及び周辺残基からの寄与をニューラルネットワークを用いて抽出することにより予測を行っている。ＭＹ法は、訓練配列の各残基位置において各アミノ酸がαヘリックスであるかの確からしさを確率分布として計算し、それからテスト配列の各領域に対し、αヘリックスの確からしさを計算する。

【０００５】

【発明が解決しようとする課題】タンパク質アミノ酸配列においては、各アミノ酸残基同士は依存関係を保持し、その局所的な立体構造や機能的な部位を形成していると考えられている。従って、タンパク質の局所的な立体構造を表現及び予測するためには、それら局所的な立体構造内の各残基間の依存関係の表現が重要であると考えられる。しかし、従来、それら残基位置間に依存する依存関係をネットワークの形で自動的に抽出する方法や、さらに、その依存関係を規則として未知データに対する予測を行う方法は、全く検討されておらず、そういった手法が確率されていなかった。

【０００６】

【課題を解決するための手段】第１の発明は、タンパク質構造の規則を学習するステップを有するタンパク質立体構造予測方法において、前記ステップが、正例と負例とからなる訓練データのアミノ酸配列を入力とし、各依存関係を条件付き確率とし、あらかじめ定められた情報量基準を用いて、該アミノ酸配列の各々の残基位置が依存している他の残基位置を、依存関係が循環しないという制限の下で決定することにより、該依存関係を表現する確率的規則を構築して出力することを特徴とする。

【０００７】第２の発明は、タンパク質構造の規則を学習するステップを有するタンパク質立体構造予測方法において、前記ステップが、正例と負例とからなる訓練データのアミノ酸配列の実数値属性を入力とし、各依存関係を条件付き確率とし、あらかじめ定められた情報量基準を用いて、該アミノ酸配列の各々の残基位置が依存している他の残基位置を、依存関係が循環しないという制限の下で決定することにより、該依存関係を表現する確率的規則を構築して出力することを特徴とする。

【０００８】

【発明の属する技術分野】次に、本発明について図面を参照して詳細に説明する。

【０００９】図１は、本発明に関連するタンパク質立体構造予測方法の実施例を説明するフローチャートである。本実施例では、対象とするタンパク質の局所的な立体構造としてαヘリックスを扱うものとする。

【００１０】ステップ１１では、αヘリックスの領域がわかっているタンパク質のアミノ酸配列に対して、同じファミリーのタンパク質、例えば、種が異なる同じタンパク質のアライメント（整合）をとり、αヘリックスに対応する部分配列を、αヘリックスの正例として抽出する。

【００１１】例えば、ヘモグロビンというタンパク質のβ鎖の場合には、ヒトのヘモグロビンのαヘリックスの位置は、Ｘ線結晶回折の結果から明らかになっており、８個のαヘリックスの領域を有することが知られている。従って、ヒトのヘモグロビンβ鎖に対して、他の種、例えば、チンパンジー、ウマなどの他の種のヘモグロビンβ鎖のアライメントを行い、８個のαヘリックスに対応する領域をαヘリックスの正例として抽出する。

【００１２】ステップ１２では、αヘリックス位置の知られているタンパク質のαヘリックスに対応する部分配列に対して、αヘリックス位置の知られているアミノ酸配列データベースの各配列のアライメントをとり、αヘリックスに対応しない部分配列を、ステップ１０で抽出されたαヘリックスの正例に対する負例として抽出する。

【００１３】ヘモグロビンβ鎖の例では、８個のαヘリックスに対応する部分配列に対して、例えば、ＰＤＢ（Ｐｒｏｔｅｉｎ　Ｄａｔａ　Ｂａｎｋ）などのタンパク質構造データベース内のいくつかの配列に対してアライメントを行い、アライメントの結果得られた各部分配列において、その配列の構造がαヘリックスではない場合に、それらを負例として抽出する。例えば、負例抽出の際のアライメントでは、一定の割合以上の相同性を保持する部分配列を負例とすることが考えられる。具体的には、アライメントによる相同性が３０％以上の部分配列を負例とする方法などがある。

【００１４】抽出するデータ数については、例えば、αヘリックスの正例となる各領域における正例と負例との割合を各領域についてそれぞれ等しくすることが考えられる。さらに、具体的には、その割合として正例、負例を同数とすることが考えられる。

【００１５】ステップ１３は、ステップ１１及びステップ１２で抽出された正例と負例から、確率的規則の実数値パラメータを計算するステップである。

【００１６】確率的規則とは、ここでは任意の与えられた配列の領域に対して、αヘリックスが対応する確率を与える確率分布のことである。各X_i　（ｉ＝１，．．．，ｎ）をそれぞれ属性値の空間として、Ｘをそれらの直積、すなわち、$X=X_1\times X_2\times\ldots\times X_n$　と書く。

【００１７】例えば、Ｘは２０種類のアミノ酸からなる一つの集合を表す場合や、また$X=X_1\times X_2$　で、X_1　が疎水性を表す数値の範囲かつX_2　が分子量を表す数値の範囲を表す場合などがある。

5

【0018】αヘリックスの正例中の長さLのウィンドウWに対し、テスト配列中の任意の長さLの領域SがW部分に対応する確からしさを以下のように求める。まず、X_t （以下、変数と呼ぶ）を配列Sの左から数えてt番目の残基位置であり、π_tを領域WにおいてX_tが依存する残基位置の集合（以下、π_t をX_t の親変数の集合、X_t をπ_t の子変数と呼ぶ）とする。ここで、

【0019】

6

＊【数1】

$$P(X_t|\pi_t)$$

【0020】を、領域Wにおいて、X_t がπ_t に依存している条件付き確率とし、領域SがW部分に対応する確からしさPw（S）は、次のように書けるものと仮定する。

【0021】

＊【数2】

$$P_W(S) = \prod_{t=1}^{L} P_W(X_t|\pi_t) \qquad (1)$$

【0022】1式の右辺は、変数をノードとし、変数間の親から子にアークを伸ばすことにより、ネットワーク構造に対応する。例えば、領域Sが3残基からなり、領域Sの各残基の結合確率が具体的に次式のように書ける※

※ものとすれば、次式は図3のネットワークに対応する。

【0023】

【数3】

$$P_W(S) = P(X_1)P(X_2|X_1)P(X_3|X_1,X_2) \qquad (2)$$

【0024】さらに、各

【0025】

【数4】

$$P(X_t|\pi_t)$$

【0026】は、与えられた正例と負例とからなる事例データから、例えば、次のようにして決定される。

【0027】まず、t番目の残基位置において、属性の実数値のとり得る範囲を重なり合わない部分領域（以下、これをセルと呼ぶ）に有限分割し、m_t を全セル数、C_iをi番目のセルとする。

【0028】t番目の位置の残基がm_t 個のセルの内のC_i に含まれる場合に、X_t の生起確率P（X_t ＝i）＝p_i （t）とする。ここで、

【0029】

【数5】

$$p_i(t) \in [0,1] \ (i = 1,\ldots,m_t)$$

★【0030】であり、これを確率パラメータと呼ぶ。図4は、有限分割の構造を示す例であるが、値が0から1の範囲をとる一つの属性により確率パラメータを推定する場合を示す。

【0031】確率パラメータは、各セルに含まれる正例及び負例のデータ数を用いて推定する。

【0032】

【数6】

$$N_i^+(t)$$

【0033】をt番目の位置でのi番目のセルに含まれる正例数、N_i （t）をt番目の位置でのi番目のセルに含まれる正例数と負例数の和とし、t番目の位置でのi番目のセルにおける推定値をp_i （t）とする。例えば、次式のラプラス推定量によって、各セルに対する確率パラメータを計算する。

【0034】

★【数7】

$$p_i(t) = \frac{N_i^+(t)+1}{N_i(t)+2} \qquad (3)$$
$$(i = 1,\ldots,m_t)$$

【0035】ただし、推定量はラプラス推定量のみならず、多くの推定量が使用できる。次に、同様に、X_t とπ_t の結合確率P（X_t ，π_t ）も推定量を用いて計算できる。例えば、π_t の要素が変数X_s のみである場合、X_s をX_t と同様に重なり合わないm_s ☆個の部分領域に有限分割し、s番目の残基がm_s 個のセル内のC_j に含まれ、

【0036】

【数8】

$$P(X_s = i, X_t = j) = p_{i,j}(t,s) \ (p_{i,j}(t,s) \in [0,1] \ (i = 1,\ldots,m_t, j = 1,\ldots m_s))$$

【0037】とし、確率パラメータ$p_{i,j}(t,s)$を推定する。

【0038】t番目，s番目の各位置において、各セルに含まれる正例及び負例のデータ数から、

【0039】

【数9】

$$N^{+}_{i,j}(t,s)$$

*【0040】を各位置においてそれぞれi，j番目のセルに含まれる正例数、$N_{i,j}(t,s)$を各位置においてそれぞれi，j番目のセルに含まれる正例数と負例数の和とする。これから、例えば、次式のラプラス推定量により、確率パラメータを推定する。

【0041】

【数10】

$$p_{i,j}(t,s)=\frac{N^{+}_{i,j}(t,s)+1}{N_{i,j}(t,s)+2} \qquad (4)$$
$$(i=1,\ldots,m_t, j=1,\ldots,m_s)$$

【0042】最後に、これら推定された確率パラメータを用いて、π_tが存在する下でのXtの条件付き確率

【0043】

【数11】

$$P(X_t|\pi_t)$$

※ように、π_tの要素が変数X_sのみであり、t番目及びs番目の各位置をそれぞれm_t個、m_s個に有限分割し、さらに、各位置の残基がセルC_i、C_jに含まれる場合、

【0045】

【数12】

【0044】を確率パラメータとして計算する。上述の※

$$P(X_t=i|X_s=j)=p_{i,j}(t|s)\ (p_{i,j}(t,s)\in[0,1](i=1,\ldots,m_t, j=1,\ldots,m_s))$$

【0046】とし、確率パラメータ

【0047】

【数13】

$$p_{i,j}(t|s)$$

★【0048】は次式のように計算する。

【0049】

【数14】

$$p_{i,j}(t|s)=\frac{p_{i,j}(t,s)}{p_j(s)} \qquad (5)$$
$$(i=1,\ldots,m_t, j=1,\ldots,m_s)$$

【0050】ステップ14では、確率的規則の構造を決定する。すなわち、各変数X_tに対し、その親変数の集合π_tを情報量基準を使用して決定するステップである。

【0051】以下、情報量基準として記述長最小（Minimum Description Lengths (MDL)）原理（以下、MDL原理）と適用した場合のネットワーク構成方法の一例について述べる。なお、MDL原理については、1978年発行の米国の雑誌「オートマティカ」（Automatica）の第14巻465－471頁記載のリサネン（Rissanen）による論文「モデリングバイ ショーテスト データ ディスクリプション」（Modeling by shortest data description）に詳しく記載されている。

【0052】MDL原理によれば、与えられた事例データから計算されるデータ記述長と規則の記述長との和が最小となる規則を最適な規則とする。従って、ステップ11、12、13において求められた正例、負例及び実数値パラメータから、ここでの確率的規則のデータ記述長及び規則の記述長を計算する。

【0053】ここで説明する例では、各残基位置ごとに、その位置と依存関係にある位置を決定していくことを考える。すなわち、各変数毎に独立にその親の変数を決定していく。

【0054】まず、t番目の残基位置に着目する。変数X_tとその親変数の集合π_tとの依存関係は、1式の確率的規則から条件付き確率

【0055】

【数15】

$$P(X_t|\pi_t)$$

【００５６】で表現される。ここで、親変数の数をｋ、親変数の残基位置を順番にt_1 からt_k とし、また、ｔ及びt_1 からt_k の残基位置での全セル数をそれぞれm_t, m_{t1} , ・・・, m_{tk} とする。さらに、変数X_t の残基がｉ、変数X_{t1} からX_{tk} の残基が、それぞれj_1 , ・・・, j_k 番目のセルに含まれるような属性を有している正例数を

【００５７】

【数１６】

$$N^{+}_{i,j_1,\cdots,j_k}(t,t_1,\cdots,t_k)$$

【００５８】、変数X_t の残基がｉ、変数X_{t1} からX_{tk} の残基が、それぞれj_1 , ・・・, j_k 番目のセルに含まれるような属性を有している正例数と負例数との和を$N_{i,j1,\cdots,jk}$ (t, t_1 , ・・ *・, t_k)、変数X_t の残基がｉ、変数X_{t1} からX_{tk} の残基が、それぞれj_1 , ・・・, j_k 番目のセルに入るような属性を有している条件付き確率の確率パラメータを

【００５９】

【数１７】

$$p_{i,j_1,\cdots,j_k}(t|t_1,\cdots,t_k)$$

【００６０】とする。

【００６１】すると、ここでの確率的規則のデータ記述長は、ステップ１３により計算された確率パラメータから、規則の対数尤度の負をとることにより、次式で与えられる。

【００６２】

【数１８】

$$-\sum_{j_k=0}^{m_{t_k}}\cdots\sum_{j_1=0}^{m_{t_1}}\sum_{i=0}^{m_t}\log\left(p_{i,j_1,\cdots,j_k}(t|t_1,\cdots,t_k)^{N^{+}_{i,j_1,\cdots,j_k}(t,t_1,\cdots,t_k)}\right.$$
$$\left.\times\left(1-p_{i,j_1,\cdots,j_k}(t|t_1,\cdots,t_k)\right)^{N_{i,j_1,\cdots,j_k}(t,t_1,\cdots,t_k)-N^{+}_{i,j_1,\cdots,j_k}(t,t_1,\cdots,t_k)}\right) \quad (6)$$

【００６３】さらに、ここでの確率的規則の規則の記述長は、次式で与えられる。

※【００６４】

※　【数１９】

$$-\sum_{j_1=0}^{m_{t_1}}\cdots\sum_{j_k=0}^{m_{t_k}}\sum_{i=0}^{m_t}\frac{\log N_{i,j_1,\cdots,j_k}(t,t_1,\cdots,t_k)}{2} \quad (7)$$

【００６５】従って、ｔ番目の残基位置に相当する変数X_t に対し、（６）式と（７）式との和を最小にするような親変数の集合を選択することにより、確率的規則の構造が決定される。

【００６６】ステップ１５では、ステップ１４において構成された確率的規則を使用し、与えられたテストデータ配列の各領域に対して、その活性度を計算する。

【００６７】ここでは、活性度として尤度を使用する。

【００６８】まず、訓練配列の正例中の任意の長さＬの領域Ｗを取り出す。このＷの各残基位置に対応する変数に対して、その親変数がステップ１４において決定されている。さらに、各変数とその親変数との依存関係を表す条件付き確率の実数値パラメータは、ステップ１３に★おいて算出されている。

【００６９】次に、この領域Ｗをテスト配列の任意の長さＬの部分配列Ｓにあてはめ、そのαヘリックスの尤度を計算する。

【００７０】例えば、Ｌ＝３の領域Ｗにおいて、（２）式のような確率的規則の構造が決定され、テスト配列のＬ＝３の領域Ｓでは、その領域内の各残基が順に、２、１、３番目のセルに入る実数値属性を有しているとする。すると、この領域ＳがＷである尤度は次式で計算できる。

【００７１】

【数２０】

$$P_W(S)=p_2(1)p_{1,2}(2|1)p_{3,2,1}(3|1,2) \quad (8)$$

【００７２】この動作を訓練配列の正例中の取り得る全ての長さＬの領域で構成された確率的規則を使用し、テスト配列中の長さＬの全ての部分配列に対して行う。

【００７３】ステップ１６では、テスト配列中の任意の長さＬの領域Ｓに対して、訓練配列中の正例の取り得る全ての長さＬの領域により算出された複数の尤度の中で、最大の尤度を選出し、領域Ｓのαヘリックスの尤度とする。

【００７４】ステップ１５及びステップ１６の動作は次のようにまとめることができる。すなわち、訓練配列中の正例中の長さＬの領域の全ての集合をＡとし、テスト配列の長さＬの部分配列Ｓに対するαヘリックスの尤度Ｐ（Ｓ）を次式により計算する。

【００７５】

【数２１】

$$P(S) = \max_{W}^{A} P_W(S) \qquad (9)$$

【0076】この動作をテスト配列の各領域に対して繰り返すことにより、テスト配列の各領域において、αヘリックスの尤度を計算する。

【0077】ここで、さらに、αヘリックス領域が複数個あれば、各αヘリックス領域について、同様な尤度計算を行ない、αヘリックス領域全体を通じて最大の尤度を最適値として選ぶ。

【0078】さらに、テスト配列内の尤度が与えられた各領域において、最大の尤度を領域内の各残基の最適値とする、あるいは、領域内の各残基に対しては、その残基を含む領域の得られた尤度の平均を各残基の最適値とする、などの方法を使用し、テストアミノ酸配列全体に対する尤度の変化を出力する。

【0079】以上の図1における学習及び予測方法は、αヘリックス以外の二次構造及びモチーフ等の局所領域の特徴抽出、さらに予測に適用できる。図2は、本発明のタンパク質立体構造予測方法の実施例を説明するフローチャートである。本実施例では、対象とする二次構造としてαヘリックスを扱うものとする。

【0080】ステップ21は、図1のステップ11と同じ処理を行ないαヘリックス領域予測のために必要な正例を抽出する。

【0081】ステップ22は、図1のステップ12と同じ処理を行ないαヘリックス領域予測のために必要な負例を抽出する。

【0082】ステップ23は、図1のステップ13と同じ処理を行ないステップ21及びステップ22で抽出された正例及び負例から、確率的規則の実数値パラメータを推定する。

【0083】ステップ24は、確率的規則の構築に制限が加えられ、局所構造領域の全変数の結合確率分布として無矛盾な確率的規則を構築するステップであり、本発明の第1の発明と第2の発明に含まれる。ここでの制限とは、確率的規則を図3のようなネットワーク構造で示した場合に、確率分布に矛盾が生じないように、アークの方向を非循環とする制限である。例えば、図3は非循環ネットワークの例であるが、この図において、X_1 からX_3 に伸びているアークを逆にX_3 からX_1 へと伸ばせば、このネットワークは循環ネットワークとなり、そのようなネットワークの生成は許さない。

【0084】制限を加える方法として、例えば、各変数に順番付けを行ない、順番の小さい変数のみを親変数として持てるとする方法、あるいは、アークに循環が生じるような依存関係が構成される場合にのみ、その依存関係を成立しないようにする方法などが考えられる。

【0085】ステップ25は、図1のステップ15と同じ処理を行ない、ステップ24を使用して構造が最適化された確率的規則を使用し、テストアミノ酸配列データの各領域に対して、その活性度を計算する。

【0086】ステップ26は、図1のステップ16と同じ処理を行ないステップ25により求められた複数の活性度から、配列全体に対する活性度の変化を出力する。

【0087】以上の図2における学習及び予測方法は、αヘリックス以外の二次構造及びモチーフ等の局所領域の特徴抽出、さらに予測に適用できる。

【0088】

【発明の効果】立体構造既知のタンパク質のアミノ酸配列情報から、局所的な立体構造さえも未知のタンパク質の局所的な立体構造を従来技術に対して高い精度で予測可能である。例えば、従来手法の一つであるMY法では、局所領域内の残基位置間の依存関係を全く考慮していなかったが、残基位置間の依存性を反映した確率的規則の構成によって、より精度の高い局所領域の特徴抽出及び予測が可能になっている。また、情報量規準に基づく最適化により、確率的規則の構造を理論的に最適化することが可能になる。

【図面の簡単な説明】

【図1】本発明に関連するタンパク質立体構造予測方法の一実施例を示すフローチャート

【図2】本発明のタンパク質立体構造予測方法の一実施例を示すフローチャート

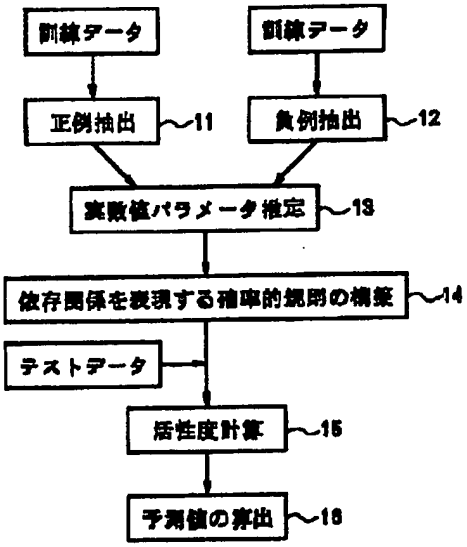
【図3】本発明で使用する確率的規則の変数間の依存関係を示す模式図。

【図4】本発明において各残基位置で行う有限分割の具体を示す模式図

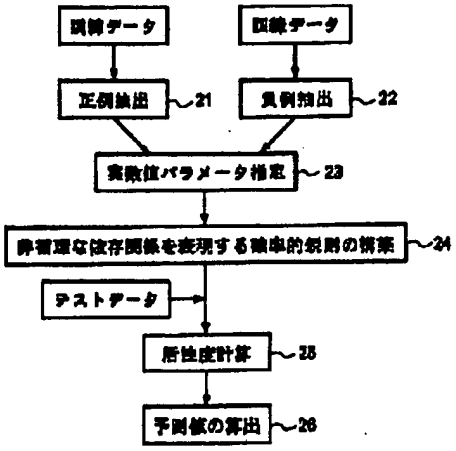
【符号の説明】

11　正例抽出
12　負例抽出
13　実数値パラメータ推定
14　確率的規則の構造の決定
15　テスト配列に対する活性度算出
16　テスト配列に対する予測値算出
21　正例抽出
22　負例抽出
23　実数値パラメータ推定
24　確率的規則の構造の決定
25　テスト配列に対する活性度計算
26　テスト配列に対する予測値算出

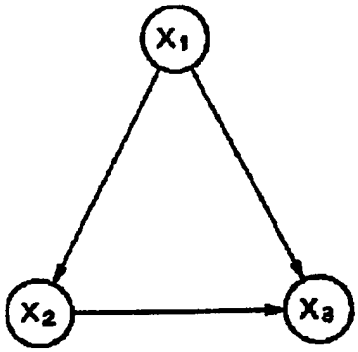
【図1】


訓練データ
訓練データ
正例抽出
11
負例抽出
12
実数値パラメータ推定
13
依存関係を表現する確率的規則の構築
14
テストデータ
活性度計算
15
予測値の算出
16

【図2】


訓練データ
訓練データ
正例抽出
21
負例抽出
22
実数値パラメータ推定
23
非循環な依存関係を表現する確率的規則の構築
24
テストデータ
活性度計算
25
予測値の算出
26

【図3】


X1
X2
X3

【図4】


1
P(Xi=1)
=Pi(ε)
C1
C2
C3
Cm
0
1